Tracking repeats using significance and transitivty.

Szklarczyk, R.; Heringa, J.

published in
Bioinformatics
2004

DOI (link to publisher)
10.1093/bioinformatics/bth911

document version
Publisher's PDF, also known as Version of record

BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i311–i317
DOI: 10.1093/bioinformatics/bth911

Tracking repeats using significance and transitivity

Radek Szklarczyk and Jaap Heringa*

Centre for Integrative Bioinformatics (IBIVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Internal repeats in coding sequences correspond to structural and functional units of proteins. Moreover, duplication of fragments of coding sequences is known to be a mechanism to facilitate evolution. Identification of repeats is crucial to shed light on the function and structure of proteins, and explain their evolutionary past. The task is difficult because during the course of evolution many repeats diverged beyond recognition.

Results: We introduce a new method TRUST, for *ab initio* determination of internal repeats in proteins. It provides an improvement in prediction quality as compared to alternative state-of-the-art methods. The increased sensitivity and accuracy of the method is achieved by exploiting the concept of transitivity of alignments. Starting from significant local suboptimal alignments, the application of transitivity allows us to (1) identify distant repeat homologues for which no alignments were found; (2) gain confidence about consistently well-aligned regions; and (3) recognize and reduce the contribution of non-homologous repeats. This re-assessment step enables us to derive a virtually noise-free profile representing a generalized repeat with high fidelity. We also obtained superior specificity by employing rigid statistical testing for self-sequence and profile-sequence alignments. Assessment was done using a database of repeat annotations based on structural superpositioning. The results show that TRUST is a useful and reliable tool for mining tandem and non-tandem repeats in protein sequence databases, capable of predicting multiple repeat types with varying intervening segments within a single sequence.

Availability: The TRUST server (together with the source code) is available at http://ibivu.cs.vu.nl/programs/trustwww

Contact: radek@cs.vu.nl

1 INTRODUCTION

Internal repeats within protein sequences have been intensely studied since they have wide-ranging implications for the evolution and function of proteins. A classical example is chymotrypsin, which evolved through the duplication of an ancestral barrel domain, such that the active site of the modern protein comprises amino acids of either domain (Heringa, 1994). Another example is the zinc finger domain, a frequent constituent of transcription factors involved in DNA binding, where the composition and copy number of individual tandem repeats confers selectivity and activity of DNA binding.

Proper delineation of repeats at the sequence level is not only important for understanding the structure and function of proteins, but is also crucial for the detection of homologous sequences and other techniques based on sequence analysis. This is because repeats often pose a problem for alignment methods that normally are ill-prepared to deal with them.

In this paper, we introduce the method TRUST (Tracking Repeats Using Significance and Transitivity), which is capable of detecting internal sequence repeats based on sequence information of an individual sequence alone. The method exploits the concept of transitivity of alignments as well as a statistical scheme optimized for the evaluation of repeat significance.

2 METHODS

2.1 Algorithm

The TRUST algorithm detects repeats without any prior knowledge. It relies on a scheme to assess the statistical significance (P-value) of repeat alignment scores, as opposed to various parameters and arbitrary thresholds used by other methods. However, the key strategy of the method is to employ transitivity: using logical inference from alignments, we introduce new information that can identify distant homologous regions and at the same time can support or contradict existing suboptimal alignments. The transitivity scheme enables us to calculate the repeat length accurately, and allows the generation of virtually noise-free and sensitive profiles.

2.1.1 Extracting alignments Detection of suboptimal alignments is performed with the Waterman–Eggert algorithm (Waterman and Eggert, 1987). In self-sequence comparison, the highest-scoring alignment trivially covers the diagonal of the dynamic-programming matrix: therefore, we mask the

*To whom correspondence should be addressed.

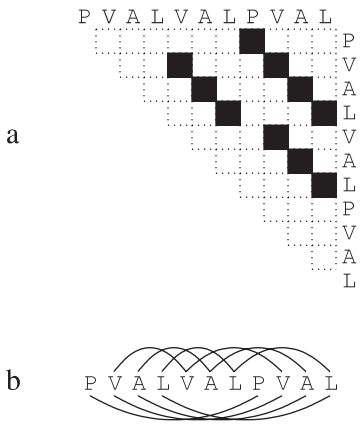

Fig. 1. (a) Matrix with the best-scoring self-alignments within the sequence PVALVALPVAL. Each black cell represents a pair of residues matched in a local alignment. The matrix diagonal and lower triangle are not shown. **(b)** Equivalent graph representation of the alignments from (a), where residues aligned are connected by edges.

matrix diagonal before the procedure starts. Note that in the self-comparison, the lower and upper triangle of the matrix are symmetrical.

An alignment can be represented as a number of dots in a two-dimensional (2D) matrix, each dot representing a matched residue pair; we call such a sequence of dots a trace (Fig. 1). A value is assigned to each trace: for traces representing alignments the value is simply the alignment score (Fig. 2a). We will use the terms 'alignment' and 'trace' interchangeably.

2.1.2 Estimating the significance of the alignments To assess the biological significance of suboptimal alignments containing repeats, we use P-values, defined as the probability of obtaining an alignment with the same score by self-alignment of scrambled sequences. Alignments with P-values lower than the default threshold of 1% are considered significant and are included in further analysis.

The distribution of the scores of highest-scoring local alignments in random sequences can be approximated with the Extreme Value Distribution (EVD) (Gumbel, 1958). When no gaps are allowed in the alignments (gap penalty $= -\infty$), the distribution of the highest alignment scores is provea to follow the EVD (Karlin and Altschul, 1990). Partial results and further empirical evidence (e.g. Waterman and Vingron, 1994a,b; Vingron and Waterman, 1994; Altschul and Gish, 1996) strongly suggests that the same distribution also applies to alignments with gaps. A benefit of the Extreme Value theory is the ease with which the distribution can be approximated, with only a limited number of scrambled sequences. We therefore determine the distributions for self-sequence alignment and profile-sequence alignment for each query sequence on the fly.

2.1.3 Transitivity Transitivity of alignments has been successfully employed in the field of sequence analysis (e.g. Notredame *et al.*, 2000) The effect of transitivity is illustrated in Figure 3. We use transitivity in the following way: if a residue i is matched with a residue j, and j is aligned to k as well, then we infer a correspondence between residues i and k (Fig. 3a). If there already exists a significant alignment containing a match between residues i and k, its validity becomes supported by the transitive alignment. In case this match did not exist between i and k, the inferred relation between i and k can affect the results when more support emerges from different alignments. In this manner, transitivity allows the detection of new alignments that were either missed or previously deemed insignificant.

Initially, each trace representing suboptimal alignment receives a value of its score. There can be up to four transitive traces generated by a pair of suboptimal alignments (Fig. 4). Therefore, the value of each new transitive trace T_{trans} is set to be one-fourth of the minimum of values of the suboptimal alignments it originated from (T_1 and T_2):

$$\text{value}(T_{\text{trans}}) = \frac{\min[\text{value}(T_1), \text{value}(T_2)]}{4}.$$

To speed up the calculations, the transitive traces are created only for suboptimal alignments, i.e. neither second-order transitive traces (transitivity applied to transitive traces) nor transitive closure (the operation repeated an infinite number of times) are calculated.

Transitive traces can overlap with suboptimal alignments obtained earlier. Therefore, we redefine the score for the match of the residues by adding the scores of all relevant traces T (transitive and non-transitive)

$$\text{score}(i, j) = \sum_{(i,j)\in T} \text{value}(T). \qquad (1)$$

Thus, the score for a residue pair that is supported by many transitive traces will be amplified.

2.1.4 Estimating the tandem repeat size If the length of the detected alignment is longer than its distance to the diagonal (Fig. 2), it is likely that a tandem repeat has been identified and that the alignment comprises a number of repeats (in the case when the number of residues matched in highest-scoring alignment is smaller than its distance to the diagonal, the length of the alignment becomes a putative repeat length). To estimate the length of a tandem repeat, we sum all scores [Equation (1)] lying at the same distance to the matrix diagonal (Fig. 2c). This process is limited to residues involved in the highest-scoring trace to avoid contributions from other types of repeats when recognizing the current type. The distance with the highest sum becomes the putative tandem repeat length L. We also include in further evaluations those distances with summed values of at least half the maximal value,

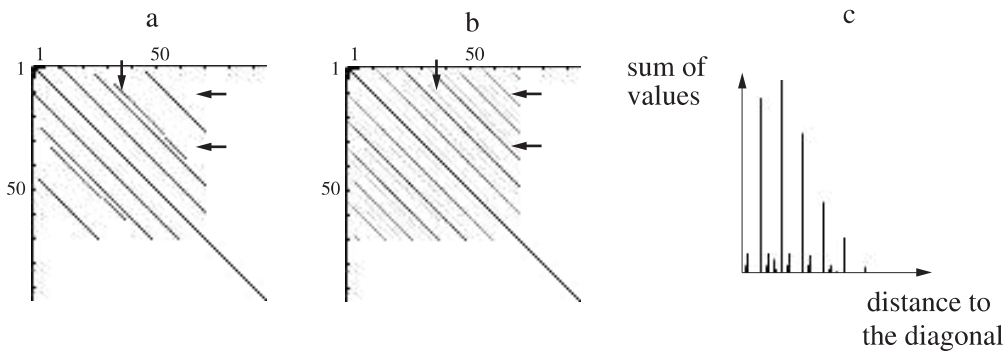

Fig. 2. Alignment matrices for a fragment of APE_SAISC-F (fragment of Apolipoprotein E of common squirrel monkey, SWISS-PROT accession number Q28995). The values of the traces are represented by grey levels. For illustrative purposes the whole matrix is shown, although the algorithm operates on the upper matrix triangle only. (**a**) Matrix showing statistically significant suboptimal alignments (vertical arrow denotes one spurious alignment); (**b**) Matrix after introducing new traces using transitivity. Suboptimal traces as in (a) and new transitive traces can be observed: previously missing traces have been reconstructed (marked with horizontal arrows) and the relative contribution of the spurious, although statistically significant alignment, is reduced (vertical arrow); (**c**) Histogram showing the sum of the scores of all traces plotted against their distance to the matrix diagonal, which is used to estimate the tandem repeat size.

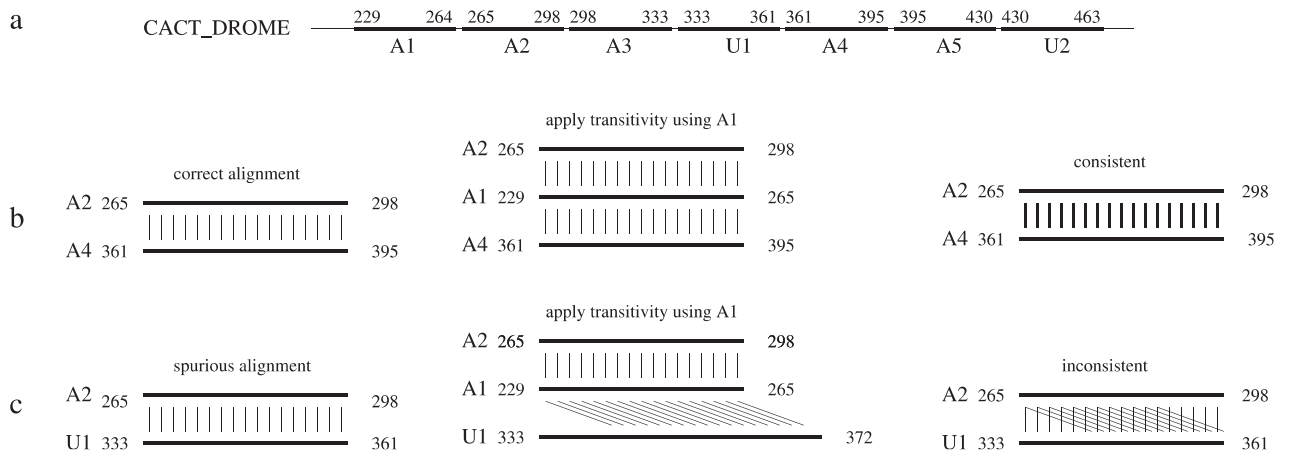

Fig. 3. Diagram explaining transitivity using CACT_DROME protein (developmental protein cactus of *Drosophila melanogaster*, SWISS-PROT accession number Q03017). (**a**) Repeats as annotated in (Bahr *et al*., 2001). (**b**) Alignments between two legitimate repeats are fully supported by transitivity, increasing their score and thereby confidence about their homology. (**c**) Alignments of repeat type A with an unknown subtype of repeat U are not confirmed when transitivity is applied.

where each of the distances selected by this procedure is also tested.

The method uses the fact that the histogram (Fig. 2c) is based on both suboptimal and transitive traces, thereby limiting the noise introduced by irrelevant local alignments that are not supported by transitivity (Fig. 2b). Furthermore, as a result of the ability to reconstruct missing traces, a precise tandem repeat size can be estimated even in cases where no suboptimal alignment at this distance from the diagonal can be found before transitivity is applied.

2.1.5 Creating the profile For the repeat length L, we compute a profile of a putative repeat set. TRUST is designed to calculate the repeat profile based on a sliding window of L consecutive columns of the trace matrix. Every residue matched in a column may participate in the profile with some weight. We calculate the weight to reflect the information from other (also transitive) traces in the neighbourhood. A weight of the residue score_w is equal to the value of its score [Equation (1)], but reduced by the value of the next highest-scoring residue no further than $L/2$ away (Fig. 5b). It can be expressed with the following equation:

$$\text{score}_w(i,j) = \max\Big[0, \text{score}(i,j) - \max_{i' \in \text{env}(i,L)} [\text{score}(i',j)]\Big],$$

where $\text{env}(i,L) = \{\lfloor i - L/2 \rfloor, \ldots, \lceil i + L/2 \rceil\} \backslash \{i\}$. The purpose of weighting is to scale down the contribution of the

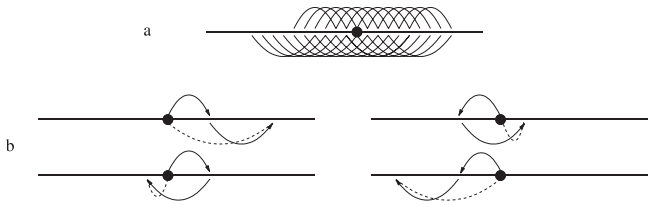

Fig. 4. View of a sequence (horizontal lines) with residue matches represented by edges. With transitivity, for each pair of suboptimal alignments, up to four new transitive traces can be created. The dot marks a single residue for which four new matches will be inferred. (**a**) Two suboptimal alignments from which a transitive trace will be created. (**b**) Arrows correspond to matches in original alignments, the directions of the arrows correspond to the direction of inference. Only matches relevant to the marked residue are depicted. New edges inferred from suboptimal alignments are shown as dashed lines.

residues aligned with less confidence (compare Fig. 5b and c), i.e. penalize residues for which traces show local alternatives.

Among all possible profiles created from L subsequent columns of the matrix of size $N \times N$, the one with the highest sum of weights is chosen. The profile starts at column $j_{\max}$ such that

$$\sum_{k=0}^{L-1} \sum_{i=1}^{N} \text{score}_w(i,\ j_{\max} + k),$$

is maximal. Having score_w and $j_{\max}$ calculated, we can obtain the relative contribution of a residue r in the k-th column of the profile

$$p_{\text{rel}}(r,k) = \sum_{i=1}^{N} \delta(s[i] = r) \cdot \text{score}_w(i, j_{\max} + k - 1),$$

where $s[i]$ is the i-th residue of the sequence, $\delta(\text{expr})$ is 1 if expr is true, and 0 otherwise. Based on p_{rel} the normalized contribution of the residue in the non-empty column is calculated with the formula

$$p(r,k) = \frac{p_{\text{rel}}(r,k)}{\sum_u p_{\text{rel}}(u,k)}.$$

Owing to TRUST's ability to recognize poorly aligned regions, this method creates a profile with minimal noise and therefore higher specificity.

2.1.6 Finding significant repeat instances After compilation of the repeat profile, a wrap-around local alignment algorithm is run (Waterman, 1995) to align the profile against the sequence and to identify the repeat instances. To infer the significance of the profile-sequence comparisons, we estimate EVD parameters by aligning the profile against shuffled sequences. The method inspects all wrap-around alignments in order to reject false positives repeat instances. The statistical significance of the single repeat instance is estimated based on the score, giving very accurate predictions also for non-integer number of repeats. In this process from all profile lengths L evaluated, the one leading to the highest number of statistically significant repeat instances is chosen.

Since many sequences contain more than one type of repeat, we iterate the above scenario to find all repeat types. This is implemented by masking the residues involved in identified repeats, and restarting the process from Section 2.1.1 (identifying self-sequence local alignments). If no statistically significant alignments can be found, iteration is terminated.

2.1.7 Implementation The program is written in Java. The time complexity is $O(N^2 + NA^2 + TLN)$ where N denotes the length of the sequence, A the number of significant suboptimal alignments, T the amount of different profile lengths investigated and L the average profile size. The execution time of the program is <1 min for a sequence of 2000 residues (Pentium III, 1.7 GHz). By default, the BLOSUM 62 substitution matrix (Henikoff and Henikoff, 1992) is used, with penalties -8 for gap opening and -2 for gap extension. To identify low-complexity regions, the seg program (Wootton and Federhen, 1993) was used.

2.2 Evaluating TRUST

The BAliBASE Benchmark Alignment Database 2.0 (Bahr *et al.*, 2001), which incorporates sets of structural repeats, was used to evaluate the quality of the TRUST method. The reference set we used contains 12 repeat families consisting of 2316 repeats in 602 sequences (up to 41 repeats per protein). The authors of BAliBASE grouped sequences in different categories (with some sequences represented in more than one category) testing different aspects of repeat detection. They also grouped the sequences into families, corresponding to repeat types.

Among many repeat types reported by the methods we tested, the type covering the most residues of reference repeats is chosen for further evaluation. A reference repeat is declared detected if at least half of its residues are covered by a single repeat reported by a program. A reported repeat detects at most one reference repeat, to avoid the situation where overestimating repeat lengths would be favoured. Reported repeats that do not detect any reference repeat reduce the accuracy of the prediction. We also calculated the number of residues overlapping with reference repeats.

To assess the repeat finding performance of TRUST, we compared it with the method RADAR (Heger and Holm, 2000). The RADAR algorithm identifies repeats based on suboptimal self-sequence alignments. The repeat boundaries are assigned such that a maximum number of integer repeats is obtained. (although a different method is used to find shorter types of repeats). RADAR additionally uses a database of precomputed multiple alignments to optimize repeat recognition. Since this database is not part of the RADAR standalone distribution, we had to use the web interface to the program, which limits the permitted query sequence length to 1000 residues

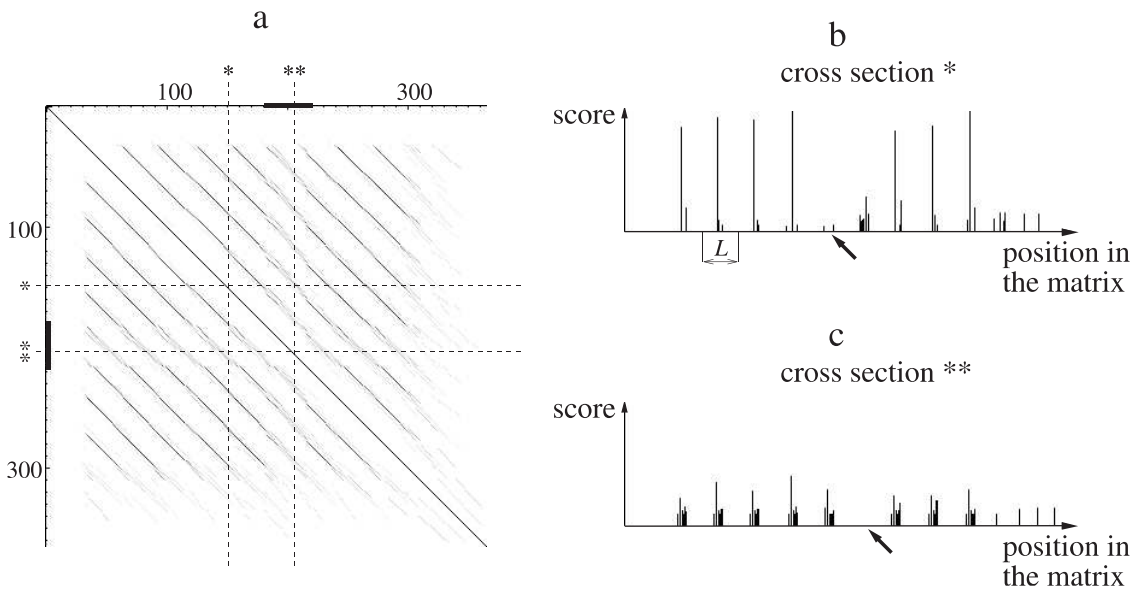

Fig. 5. Cross sections through the transitive trace matrix at marked positions. (**a**) The matrix of all traces [Equation (1)]. The thick line on the axes denotes one divergent repeat. (**b**) and (**c**) Values in the columns of the matrix. *L* denotes the length of the tandem repeat, the intersection with the diagonal is denoted with the arrow.

(this limited the number of BAliBASE sequences from 602 to 530).

The sensitivity and accuracy of repeat detection was assessed using manually curated annotations of BAliBASE, which are based on structural superpositions (Bahr *et al*., 2001). Sensitivity of detecting repeats is measured as the ratio of the number of repeats detected to the number of repeats annotated in BAliBASE. Accuracy is defined as the ratio of the number of repeats properly detected to the total number of reported repeats. The same notions apply to the residues involved in the repeats.

3 RESULTS

3.1 Sensitivity and accuracy

Compared with the RADAR program, TRUST shows high sensitivity across the range of categories (Table 1), both for repeats (72% in comparison with 64%) and residues. TRUST also shows a higher accuracy of repeat prediction, although the accuracy is the same as for RADAR if repeat residues are counted.

The quality of the prediction of the repeat length was much higher for the TRUST method. The difference between the estimated repeat length and the median of the length from BAliBASE is no more than 1 residue for 44% of repeat types (19% in case of RADAR). If we count the number of the repeats with their length predicted properly within 10% margin of their reference length, TRUST correctly reports 65% of repeat types (40% for RADAR).

The sensitivity and accuracy for repeat detection was also measured for BAliBASE repeat families. The sensitivity of repeat detection of the TRUST method was 69% (64% for RADAR) with an accuracy of 92% (78% for RADAR). For example, one of the most challenging families in BAliBASE is Myb DNA-binding domain (the family consist of 138 proteins, with repeats around 50 residues long). Two factors render detection of repeats within this family difficult: (1) scarcity of repeats (most proteins have only two) and (2) their divergence. For this family of proteins, RADAR detects considerably more repeats (63%) than TRUST (43%) when run with default parameters (Table 2), although at a price of a much lower accuracy. This suggests that many of the repeats found by RADAR have low statistical significance, and close inspection confirms that many repeats are inferred from insignificant alignments. In TRUST, to provide the user with the possibility to find repeats that would otherwise be discarded based on low statistical score, the '-force' program parameter can be used (also available via the web interface). This option forces the TRUST program to include the highest-scoring suboptimal alignment in calculations, even if the statistical significance is below the default threshold. This feature should be used only if the user is convinced that the sequence contains internal repeats. For Myb DNA-binding repeat family using this option increases the sensitivity, at the rate of 10 false positive repeats for this family.

3.2 Specificity

When repeat detection is automated, e.g. in large-scale database mining, or used as a filtering step before multiple

Table 1. Sensitivity and accuracy for different categories of BAliBASE

Category	Repeats Sensitivity (%) TRUST	RADAR	Accuracy (%) TRUST	RADAR	Residues Sensitivity (%) TRUST	RADAR	Accuracy (%) TRUST	RADAR
1a	90	93	97	91	89	84	94	92
1b	96	68	89	67	91	83	76	78
2a	56	55	91	68	61	65	88	85
2b	73	67	92	85	83	77	86	85
2c	74	65	94	82	85	77	86	85
3	85	62	96	63	93	81	88	87
4	69	63	72	59	73	71	45	57
All repeats (2799)	72	64	91	75	81	76	83	83
All categories (6)	78	68	90	74	82	77	80	81

The better result of the two methods is underlined. 'All repeats' row is the average sensitivity with equal contribution of every repeat, 'All categories' row is sensitivity with equal contribution of every category. The difference in the number of detected repeats in the category 1a is one repeat, and in 2a it is six repeats.

Table 2. The Results of predictions for Myb repeat family with 249 repeats

Method	Sensitivity (%)	Accuracy (%)	False positives
TRUST	43	82	24
TRUST -force	75	85	34
RADAR	63	64	86

'-force' is a parameter of TRUST, described in the text. Despite scarcity and the divergence of the repeats within this family, TRUST predictions are very accurate even with increased sensitivity.

alignment, there should be no prior assumption whether the sequence contains any repeats (otherwise many false positive repeats will be reported). This property (specificity) was tested using 100 random sequences, with a reasonable expectation of no repeats to be reported. The random sequences were generated using the residue composition of the SWISS-PROT database (Boeckmann *et al*., 2003), each sequence 1000 residues long. TRUST did not detect any significant repeat occurrences in the sequences, in comparison to an average of five repeats per sequence reported by RADAR (with size ranging from 10 to 200 residues, the median was 34 residues).

4 DISCUSSION

4.1 Related work

In addition to TRUST and RADAR another repeat detection tool that we considered is REPRO (Heringa and Argos, 1993), which has already been used for more than a decade to find repeats in proteins. It is quite a sensitive algorithm that is capable of finding both tandemly and distantly located repeats, as well as repeats of different types. Recently, an equivalent but faster algorithm has been proposed (Romein *et al*., 2003, http://www.sc-conference.org/sc2003/paperpdfs/pap189.pdf). Since the program parameters must be established by trial and error for each query sequence, we could not use it in an automated way.

Pellegrini *et al*. (1999) present a tool to analyze self-sequence alignments. The program reports the consensus size of the repeat, the number of repeat occurrences and a set of suboptimal overlapping self-alignments, which would have been used to infer the repeats from. Although the method has been used to derive a general global census of repeats (Marcotte *et al*., 1999), the lack of repeat boundary identification prevents its use in a comparison, such as carried out here between RADAR and TRUST.

Andrade *et al*. (2000) devised an iterative algorithm based on score distributions from profile analysis. The method allows the detection of 11 currently implemented repeat families, and therefore could not be included in our evaluation. Nonetheless, the method was used to find thousands of previously unrecognized repeat instances, while suggestions were made to merge several repeat families that previously were thought to be distinct.

4.2 Conclusions

Statistically significant alignments may contain non-homologous fragments, which are aligned only because they are surrounded by parts of high similarity. This can happen when a series of tandem repeats quickly diverges beyond recognition, possibly after losing its original structure or function (Fig. 5, the thick line on the axes denotes one divergent repeat). In the absence of biological evidence of homology, such fragments would add noise to the repeat profile, and if their match against the profile does not lead to statistically significant scores, they should not be reported as part of the repeat family. In the TRUST method, such divergent regions can easily be noted in the matrix of transitive traces, because they typically contain many gaps when aligned with legitimate repeats. Furthermore, the pattern of gaps will not be consistent

among different alignments, such that these will hardly be supported by transitive traces (Fig. 5b and c illustrate the aligned residues supported and unsupported by transitivity).

Introducing fixed thresholds increases the danger of producing false negatives on one side, or false positives on the other, especially when the thresholds are not dependent on the exchange matrix, sequence size or residue composition. Therefore we relied strongly on statistical significance wherever possible, resulting in the high specificity of our tool. Sensitivity and correctness of repeat size calculation and boundary prediction is achieved through transitivity, allowing us at the same time to use simple profile creation protocols. By exploring the concept of transitivity, missing traces are reconstructed, and the relative role of spurious ones reduced. Therefore, with profiles based on reconstructed traces we can find many repeat occurrences without sacrificing accuracy.

ACKNOWLEDGEMENTS

The authors would like to thank the following people: Jose Ferreira for discussions about the extreme value theory and Katarzyna Gryszczuk, Jens Kleinjung, John Romein and Michael Sammeth for suggestions leading to improvement of the manuscript.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.* **266**, 460–480.

Andrade,M.A., Ponting,C.P., Gibson,T.J. and Bork,P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.*, **298**, 521–537.

Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York.

Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Prot. Struct. Funct. Genet.*, **41**, 224–237.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Heringa,J. (1994) The evolution and recognition of protein sequence repeats. *Comput. Chem.*, **18**, 233–243.

Heringa,J. and Argos,P. (1993) A method to recognize distant repeats in protein sequences. *Prot. Struct. Funct. Genet.*, **17**, 391–411.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1999) a census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Pellegrini,M., Marcotte,E.M. and Yeates,T.O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Prot. Struct. Funct. Genet.*, **35**, 440–446.

Romein,J., Heringa,J. and Bal,H. (2003) A million-fold speed improvement in genomic repeats detection. In *SuperComputing'03*, Phoenix, AZ.

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.*, **235**, 1–12.

Waterman,M.S. (1995) *Introduction to Computational Biology*. Chapman & Hall, London, UK.

Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to TRNA–RRNA comparisons. *J. Mol. Biol.*, **197**, 723–728.

Waterman,M.S. and Vingron,M. (1994a) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Waterman,M.S. and Vingron,M. (1994b) Sequence comparison significance and poisson approximation. *Stat. Sci.*, **9**, 367–381.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.